Is the Property of Being Positively Correlated Transitive?

Eric LANGFORD, Neil SCHWERTMAN, and Margaret OWENS

Suppose that X, Y, and Z are random variables and that X and Y are positively correlated and that Y and Z are likewise positively correlated. Does it follow that X and Z must be positively correlated? As we shall see by example, the answer is (perhaps surprisingly) "no." We prove, though, that if the correlations are sufficiently close to 1, then X and Z must be positively correlated. We also prove a general inequality that relates the three correlations. The ideas should be accessible to students in a first (postcalculus) course in probability and statistics.

KEY WORDS: Correlation.

1. INTRODUCTION

Suppose that X, Y, and Z are random variables and that X and Y are positively correlated and that Y and Z are likewise positively correlated. It seems reasonable to assume that X and Z must also be positively correlated, since X and Y being positively correlated means that "as X increases, Y tends to increase," and Y and Z being positively correlated means that "as Y increases, Z tends to increase." One would think that this would imply that as X increases, Z would tend to increase also. One might even attempt a proof along the lines of noting that the regression line of Y on X is of the form $y = a_1x + b_1$ and the regression line of Z on Y is of the form $z = a_2y + b_2$, where a_1 and a_2 are both positive. It should follow that the regression line of Z on X should be "close" to $z = a_2(a_1x + b_1) + b_2 = (a_2a_1)x + (a_2b_1 + b_2)$, where a_2a_1 is positive, implying positive correlation between X and Z. The only trouble with this "proof" is that it is false! We now give an example that is easily accessible to students in a beginning course in statistics.

2. EXAMPLE 1

Three people work in an office: Mr. A, Ms. B, and Mr. C. Mr. A is 6 feet tall, weighs 190 pounds and is 45 years of age. Ms. B is 5 feet, 8 inches tall, weighs 180 pounds, and is 50 years of age. Mr. C is 5 feet, 10 inches tall, weighs 170 pounds, and is 40 years of age. If we let X, Y, and Z, respectively, denote the height, weight, and age of each of the three people, it is easily computed that the correlations are given by $\rho_{XY} = \rho_{YZ} = 1/2$ and that $\rho_{XZ} = -1/2$.

Eric Langford is Professor, Neil Schwertman is Professor, and Margaret Owens is Professor, California State University, Chico. The authors thank Rebecca Langford, an undergraduate student, for her assistance in drawing Figure 1. We also thank the editor and two referees for their helpful suggestions.

3. EXAMPLE 2

Obviously, the above example is artificial. Table 1 is an example of the same situation from real-life data. The data are for all of the New York Yankees with at least 300 at-bats at the end of the 2000 regular season. The numbers X, Y, and Z represent the number of triples, base hits, and home runs for that particular player.

If we calculate the correlations in Table 1, we find that $\rho_{XY} = 0.526$, $\rho_{YZ} = 0.293$, whereas $\rho_{XZ} = -0.096$. It is easy to explain these correlations after the fact: certainly one would expect that the more hits a player has, the more triples and the more home runs he would have. On the other hand, it is not surprising that the number of triples and the number of home runs should be negatively correlated, as smaller, faster players (e.g., Polonia) tend to hit triples and not home runs, whereas larger, more powerful players (e.g., Canseco) tend to hit home runs and not triples.

It is clear that if X, Y, and Z are random variables with the above correlation properties (i.e., that ρ_{XY} and ρ_{YZ} are positive and that ρ_{XZ} is negative), then the random variables X, $-Y$, and Z will be mutually negatively correlated, and vice versa. Hence an example of one phenomenon automatically produces an example of the other. But if X, Y, and Z are any random variables such that

$$\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y} + \frac{Z}{\sigma_Z} = \text{constant}$$

they will be mutually negatively correlated. For example, roll a die: if the die shows 1 or 2, then $X = 1$ and $Y = Z = 0$; if the die shows 3 or 4, then $Y = 1$ and $X = Z = 0$; if the die shows 5 or 6, then $Z = 1$ and $X = Y = 0$. Or choose a point at random in the triangular piece of the plane $x + y + z = 1$ which lies in the first octant. Then let X, Y, and Z denote the coordinates of the chosen point.

We thank a referee for pointing out the following theorem which provides an inexhaustible supply of examples:

Theorem 1. Let U, V, and W be any nontrivial independent random variables and define X, Y, and Z by $X = U + V$, $Y = W + V$, and $Z = W - U$. Then ρ_{XY} and ρ_{YZ} are positive and ρ_{XZ} is negative.

Table 1. 2000 New York Yankees With at Least 300 At-Bats

Player	X	Y	Z
Jeter	4	201	15
Williams	6	165	30
Posada	1	145	28
Justice	1	150	41
O'Neill	0	160	18
Knoblauch	2	113	5
Polonia	5	140	7
Martinez	4	147	16
Canseco	0	83	15
Brosius	0	108	16

4. ANALYSIS

Let us consider first the case of three simple random variables X, Y, and Z. (By "simple" we mean that each assumes only finitely many distinct values.) Technically, X, Y, and Z are all functions defined on a common finite set $\Omega = \{a_1, a_2, \ldots, a_n\}$. We can now identify each random variable (function) with a vector in n-dimensional Euclidean space via the correspondence

$$X \leftrightarrow \mathbf{X} = (X(a_1), X(a_2), \ldots, X(a_n)),$$

and similarly for Y and Z. If we assume that Ω is equipped with a uniform probability measure (as will be the case when we are dealing with "data"), then the correlations of the random variables X, Y, and Z are reflected in the geometric properties of the corresponding vectors $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$. Since variances and covariances are translation-invariant, we can assume without loss of generality that $E(X) = E(Y) = E(Z) = 0$.

From vector analysis, we recall that the dot product (scalar product) of two vectors $\mathbf{X}$ and $\mathbf{Y}$ in n-dimensional Euclidean space is given by the formula $\mathbf{X} \cdot \mathbf{Y} = |\mathbf{X}||\mathbf{Y}| \cos\theta_{XY}$, where θ_{XY} is the angle between $\mathbf{X}$ and $\mathbf{Y}$. Since $\mathbf{X} \cdot \mathbf{Y} = \sum x_i y_i$ and since $|\mathbf{X}|^2 = \mathbf{X} \cdot \mathbf{X}$ it follows that

$$\cos\theta_{XY} = \frac{\mathbf{X} \cdot \mathbf{Y}}{|\mathbf{X}||\mathbf{Y}|} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} = \rho_{XY},$$

so that the correlation ρ_{XY} of the random variables X and Y is simply the cosine of the angle between the two corresponding vectors $\mathbf{X}$ and $\mathbf{Y}$! Even though we may be in n-dimensional space, everything takes place in a three-dimensional subspace (namely the subspace spanned by $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$), so that we can still use geometric arguments. Hence two random variables have positive correlation if the angle between their corresponding vectors is acute, negative correlation if the angle between their corresponding vectors is obtuse, and zero correlation if the angle between their corresponding vectors is a right angle. If the angle is zero, the random variables have correlation 1 and if the angle is a straight angle, the random variables have correlation -1. (This will be true of all random variables when "angle" is interpreted in an abstract sense. See the Appendix.)

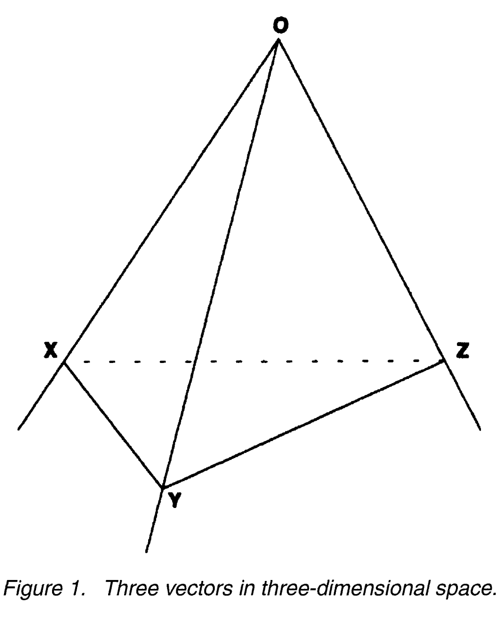

Figure 1. Three vectors in three-dimensional space.

If we apply this analysis to Example 1, we can naturally take $\Omega = \{A, B, C\}$ so that letting X denote height in inches (after subtracting off the mean 70), Y denote weight in pounds (after subtracting off the mean 180), and Z denote age in years (after subtracting off the mean 45) we see that $X(A) = 2, X(B) = -2$, and $X(C) = 0$ and similarly for Y and Z. Thus, the corresponding vectors are $\mathbf{X} = (2, -2, 0)$, $\mathbf{Y} = (10, 0, -10)$, and $\mathbf{Z} = (0, 5, -5)$ and it follows that the angle between $\mathbf{X}$ and $\mathbf{Y}$ and the angle between $\mathbf{Y}$ and $\mathbf{Z}$ are both equal to 60°, whereas the angle between $\mathbf{X}$ and $\mathbf{Z}$ is 120°.

It should be obvious geometrically that if the angles between $\mathbf{X}$ and $\mathbf{Y}$ and $\mathbf{Y}$ and $\mathbf{Z}$ are small enough (i.e., if the correlations between X and Y and Y and Z are large enough), then the angle between $\mathbf{X}$ and $\mathbf{Z}$ will be an acute angle and so X and Z must be positively correlated. Details are given in the following.

Theorem 2. Suppose that X, Y, Z are random variables and that X and Y, and Y and Z, are positively correlated with correlations ρ_{XY} and ρ_{YZ}, respectively. Suppose further that $\rho_{XY}^2 + \rho_{YZ}^2 > 1$. Then X and Z are positively correlated.

Proof. We assume for purposes of the geometrical proof that the random variables are simple and that Ω is equipped with a uniform probability measure. [We include an algebraic proof in the appendix (Theorem A.1) which depends only on the properties of the inner product in a Hilbert space and therefore applies to all random variables.]

Assume that $E(X) = E(Y) = E(Z) = 0$ and consider the corresponding vectors in n-dimensional space as before. It is known that the angle θ_{XZ} between $\mathbf{X}$ and $\mathbf{Z}$ (i.e., $\angle$ XOZ) is less than or equal to the sum of the angle θ_{XY} between $\mathbf{X}$ and $\mathbf{Y}$ (i.e., $\angle$XOY) and the angle θ_{YZ} between $\mathbf{Y}$ and $\mathbf{Z}$ (i.e., $\angle$YOZ). [See Sigley and Stratton (1942), Theorem 35, p. 46 and refer to Figure 1.]

Since these three angles are given by $\cos^{-1}\rho_{XZ}$, $\cos^{-1}\rho_{XY}$, and $\cos^{-1}\rho_{YZ}$, respectively, and since $\cos^{-1}\rho_{XY}$ and $\cos^{-1}\rho_{YZ}$ are acute because ρ_{XY} and ρ_{YZ} are positive, it follows that ρ_{XZ} will be positive whenever $\cos^{-1}\rho_{XY} + \cos^{-1}\rho_{YZ}$ is acute, that is, whenever

$$\cos\left(\cos^{-1}\rho_{XY} + \cos^{-1}\rho_{YZ}\right) > 0.$$

Expanding the left-hand side out by the addition formula for the cosine, we see that this is equivalent to

$$\cos\left(\cos^{-1}\rho_{XY}\right)\cos\left(\cos^{-1}\rho_{YZ}\right) - \sin\left(\cos^{-1}\rho_{XY}\right)\sin\left(\cos^{-1}\rho_{YZ}\right) > 0$$

$$\rho_{XY}\rho_{YZ} - \sqrt{\left(1-\rho_{XY}^2\right)\left(1-\rho_{YZ}^2\right)} > 0.$$

Rearranging, squaring both sides, and then simplifying yields the equivalent condition

$$\rho_{XY}^2 + \rho_{YZ}^2 > 1.$$

A similar analysis can be used to show the following inequality. Notice that Theorem 2 is then an immediate consequence of Theorem 3.

Theorem 3. For any three random variables, the correlations must satisfy the following inequality:

$$\rho_{XY}\rho_{YZ} - \sqrt{(1-\rho_{XY}^2)(1-\rho_{YZ}^2)} \le \rho_{XZ} \le \rho_{XY}\rho_{YZ} + \sqrt{(1-\rho_{XY}^2)(1-\rho_{YZ}^2)},$$

which can be rewritten in the more symmetric form

$$\rho_{XY}^2 + \rho_{YZ}^2 + \rho_{XZ}^2 - 1 \le 2\rho_{XY}\rho_{YZ}\rho_{XZ}.$$

Proof. We note that this last inequality can be found as Problem 15.2 (p. 536) of Stuart and Ord (1994), and the one immediately before can be found as Problem 15.3 (p. 387) of Kendall (1948). We thank the editor for pointing this out to us.

Since we are giving a geometric proof, we make the same assumptions as in Theorem 2 and we again note that Theorem A.1 in the appendix proves the result for any random variables whatsoever.

By rearranging the inequality of Sigley and Stratton (1942), we can show that

$$|\theta_{XY} - \theta_{YZ}| \le \theta_{XZ} \le \theta_{XY} + \theta_{YZ}. \quad (1)$$

Assuming for the moment that the right-hand side of (1) is less than a straight angle, we can apply the cosine function to each part of (1), and since the cosine is decreasing on $[0, \pi]$, the inequalities are reversed and we have that

$$\cos(\theta_{XY} - \theta_{YZ}) \ge \cos\theta_{XZ} \ge \cos(\theta_{XY} + \theta_{YZ})$$

using the fact that the cosine is an even function so we can remove the absolute value bars. As in the proof of the theorem, it follows that

$$\rho_{XY}\rho_{YZ} - \sqrt{(1-\rho_{XY}^2)(1-\rho_{YZ}^2)} \le \rho_{XZ} \le \rho_{XY}\rho_{YZ} + \sqrt{(1-\rho_{XY}^2)(1-\rho_{YZ}^2)},$$

which becomes, on rearranging

$$|\rho_{XZ} - \rho_{XY}\rho_{YZ}| \le \sqrt{(1-\rho_{XY}^2)(1-\rho_{YZ}^2)}.$$

Now, squaring both sides and simplifying, we get the alternative form. If, however, the right-hand side of (1) lies in the interval $[\pi, 2\pi]$, we simply note that because $\theta_{XY}+\theta_{YZ}+\theta_{XZ} \le 2\pi$, we have $\theta_{XZ} \le 2\pi - \theta_{XY} - \theta_{YZ}$ and since the right-hand side of this lies in the interval $[0, \pi]$ we can say that

$$\cos\theta_{XZ} \ge \cos(2\pi - \theta_{XY} - \theta_{YZ}) = \cos(\theta_{XY} + \theta_{YZ})$$

and the proof follows as before.

An alternative proof of Theorem 3 can be given. Because of the result on page 511 of Stuart and Ord (1994), this is probably the proof of their Problem 15.2 that is intended in that reference.

Alternative proof. As before, X, Y, Z can be identified with vectors $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, only this time we shall think of them as n-dimensional column vectors. (Recall that we are still assuming that $E(X) = E(Y) = E(Z) = 0$.) Consider the $n \times 3$ matrix A whose columns are respectively $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$. Then

$$M = \frac{1}{n}A^T A$$

is simply the variance-covariance matrix of X, Y, Z

$$M = \begin{bmatrix} \operatorname{var}(X) & \operatorname{cov}(X,Y) & \operatorname{cov}(X,Z) \\ \operatorname{cov}(X,Y) & \operatorname{var}(Y) & \operatorname{cov}(Y,Z) \\ \operatorname{cov}(X,Z) & \operatorname{cov}(Y,Z) & \operatorname{var}(Z) \end{bmatrix}.$$

Dividing the first row and the first column by $\sigma_X = \sqrt{\operatorname{var}(X)}$ and similarly for the other two rows and columns we see that

$$\det M = \operatorname{var}(X)\operatorname{var}(Y)\operatorname{var}(Z)\det\begin{bmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho_{YZ} \\ \rho_{XZ} & \rho_{YZ} & 1 \end{bmatrix}.$$

But it is known (see Graybill 1961, p. 4) that M is positive semidefinite so that its determinant must be nonnegative; that is,

$$\det\begin{bmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho_{YZ} \\ \rho_{XZ} & \rho_{YZ} & 1 \end{bmatrix} \ge 0.$$

Calculating the determinant gives

$$1 - \rho_{XY}^2 - \rho_{YZ}^2 - \rho_{XZ}^2 + 2\,\rho_{XY}\,\rho_{YZ}\,\rho_{XZ} \ge 0$$

which is the desired result.

APPENDIX

Let $(\Omega, \mathcal{A}, P)$ be a probability space and consider the space of all random variables (i.e., measurable functions on Ω) with expectation 0 which have finite variance. This space is a Hilbert space under the inner product

$$(X, Y) = \int_\Omega X(\omega)Y(\omega)dP(\omega).$$

Moreover $\operatorname{var}(X) = (X, X), \sigma_X = \sqrt{(X,X)} = |X|$ and $\rho_{XY} = \frac{(X,Y)}{|X||Y|}$.

The most general form of Theorem 3 now follows from the following theorem.

Theorem A.1. Let $\mathcal{H}$ be any Hilbert space and suppose that x, y, z are elements of $\mathcal{H}$. Assume without loss of generality that $|x|, |y|, |z|$ are all 1. Then

$$1 - (x,y)^2 - (y,z)^2 - (x,z)^2 + 2(x,y)(y,z)(x,z) \ge 0.$$

Proof. If $(y,z)^2 = 1$ (i.e., if $(y,z) = \pm 1$) then either $y = z$ or $y = -z$, and in either case, the inequality is trivially an equality. Suppose then that $(y,z)^2 < 1$. (Evidently $(y,z)^2 \le 1$ by the Schwarz Inequality.) Now project x onto the subspace spanned by y and z to get Px. This can be done by using the Gram–Schmidt procedure or by using calculus methods to find λ, μ that minimize

$$|x - (\lambda y + \mu z)|^2 = (x - \lambda y - \mu z, x - \lambda y - \mu z).$$

[This uses the characterization of Px as the closest point in the subspace to x thus making the residual $x - Px$ perpendicular to

Px (see Berberian 1961, pp. 74–75).] Either method yields the formula

$$Px = \frac{[(x,y) - (x,z)(y,z)]\,y + [(x,y) - (x,y)(y,z)]\,z}{1 - (y,z)^2}.$$

Now look at $|x - Px|^2 = (x - Px, x - Px) = (x, x - Px) - (Px, x - Px) = (x, x - Px) - 0 = (x,x) - (x, Px) = 1 - (x, Px)$. Since $|x - Px|^2 \geq 0$, we have that $0 \leq 1 - (x, Px)$. Now expanding out the inner product (x, Px) we have the desired result.

[Received May 2000. Revised June 2001.]

REFERENCES

Berberian, S. K. (1961), *Introduction to Hilbert Space*, New York: Oxford University Press.

Graybill, F. A. (1961), *An Introduction to Linear Statistical Models* (vol. I), New York: McGraw Hill.

Kendall, M. G. (1948), *The Advanced Theory of Statistics* (vol. I), (4th ed.), London: Charles Griffin.

Sigley, D. T., and Stratton, W. T. (1942), *Solid Geometry and Mensuration*, New York: Dryden Press, Inc.

Stuart, A., and Ord, J. K. (1994), *Kendall's Advanced Theory of Statistics* (vol. I) (6th ed.), London: Edward Arnold.